Exploring Adversarial Robustness of Multi-sensor Perception Systems in Self Driving - Supplementary

[1,2] **James Tu** [3] **Huichen Li** **Xinchen Yan** [1,2] **Mengye Ren** [1,2] **Yun Chen**
[1] **Ming Liang** [4] **Eilyan Bitar** **Ersin Yumer** [1,2] **Raquel Urtasun**
Waabi[1], University of Toronto[2], UIUC[3], Cornell University[4]
{jtu, mren, yun, urtasun}@cs.toronto.edu
huichen3@illinois.edu eyb5@cornell.edu
{skywalkeryxc,liangming.elgoog,meyumer}@gmail.com

1 Additional Implementation Details

Rooftop Approximation We convert all the vehicle meshes from our vehicle bank to to a set of SDF volumes, which are then projected into a lower dimensional latent code of length 32 with PCA. Then for each point cloud $P = \{p : p \in \mathbb{R}^3\}$ which we wish to fit a mesh to, we optimize the latent code z^* such that

$$z^* = \underset{z}{\operatorname{argmin}} \sum_{p \in P} H(p; z)^2. \tag{1}$$

Here $H(\cdot\,; z)$ is the SDF volume decoded from z and z^* minimizes the mean squared distance from each point in P to the zero-level surface of the SDF. To learn z^* we initialize with $z = \mathbf{0}$ and perform 200 steps of gradient updates using Adam with learning rate 0.001. Furthermore, the point cloud is also transformed to a canonical coordinate frame. While we attempt this approximation for all point clouds, a good fit is not always achievable since distant or occluded point clouds have a limited number of points. Thus, if the mean squared distance after fitting is greater than $0.02m^2$, then we default to using the top of the bounding box annotation as the roof.

Detection Model For the KITTI dataset we directly adopt the MMF model [1]. The multi-sensor detector has two separate branches to extract features from images and LiDAR, in which LiDAR points are voxelized and processed as a BEV image with 2D convolutions while images are processed with a ResNet backbone [2]. For sensor fusion, each voxel in BEV performs K-nearest neighbour (KNN) to sample close-by LiDAR points. We follow the orginal implementation and set K=1 in our fusion modules. We use a slightly different variation of the MMF model for XENITH that is tuned for more complex scenarios and faster inference. The differences can be summarized as follows:

1. The refinement module is removed as two-stage detection is slow in practice.
2. We modify the LiDAR feature extraction network from residual blocks to a feature pyramid network [3] style to enable cross-scale fusion of features at different resolutions.
3. The number of image-LiDAR fusion modules is s reduced from 4 to 1, only fusing at a single feature resolution.

Additional Dataset Details Both datasets were captured with 64 beam LiDARs spinning at 10Hz. KITTI images are captured at a 370×1224 resolution and XENITH images are captured at a 320×2048 resolution. We filtered out samples from the dataset according to the following criteria:

1. Image is captured during the day, between 9 A.M. and 5 P.M. so that objects have sufficient lighting.
2. Vehicle sample is within 70m longitudinally from the ego and 40m laterally.
3. After simulation, at least 70% of the pixels of the inserted mesh is visible in the image.

Depth Completion Our depth completion model adopts the architecture introduced in [4] and we initialize the model with pre-trained weights from COCO. The model takes RGB images concatenated with a sparse depth image obtained from projecting LiDAR points onto the image

5th Conference on Robot Learning (CoRL 2021), London, UK.

	FN ASR	FP ASR	ASR	AP (clean)
K = 0	23.80%	10.70%	32.60%	83.02
K = 1	43.15%	11.77%	49.76%	84.64
K = 3	25.80%	13.35%	37.36%	84.50
K = 5	36.67%	7.13%	41.54%	84.44
K = 7	59.25%	7.55%	62.05%	84.74

Table 1: We sweep K, the number of LiDAR points used to query image features for each BEV voxel. Overall, there is no clear trend between K and robustness.

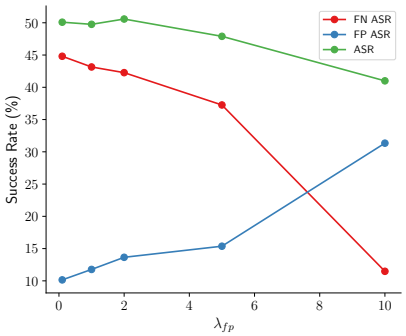

Figure 1: Sweep of λ_{fp}, the term which weights the false positive component of the loss term. Increasing λ_{fp} trade false negative success for false positives.

plane. For training, we use the official depth labels for KITTI and use aggregated LiDAR depths for XENITH for supervision. We adopt the training loss and schedule from [5].

2 Additional Experiments

λ_{fp} Sweep In our experiments we set $\lambda_{fp} = 1$, here we vary the weighting coefficient to analyze how the adversary adapts to focus more on either generating false negatives or false positives. We sweep λ_{fp} and show the results in Figure 1. At $\lambda_{fp} = 10$, the attack can trade most of the false negative success for false positives. In our main experiments, we chose to focus more on false negatives as missed detections is far more problematic.

KNN Following previous implementations of multi-sensor fusion [1], when each BEV voxel perform K-Nearest Neighbor search to query for LiDAR points, we set $K = 1$ in our main experiments. Here, we vary K and conduct our attacks on models that are retrained to use more LiDAR points for each BEV voxel during fusion. Results are shown in Table 1. Note that $K = 0$ is the equivalent of a LiDAR only model. Overall there is no clear trend between this parameter and robustness to our attack.

Success Rate Visualization To better stand where the attack is strong, we visualize the attack success rate across location in BEV. The visualization is shown in Figure 2. Host vehicles that are farther away are much easier to attack than those close by.

Qualitative Examples Additional qualitative examples are shown in Figure 3 for XENITH and Figure 4 for KITTI. Furthermore, please see our supplementary video *supplementary_video.mp4* for video demonstrations.

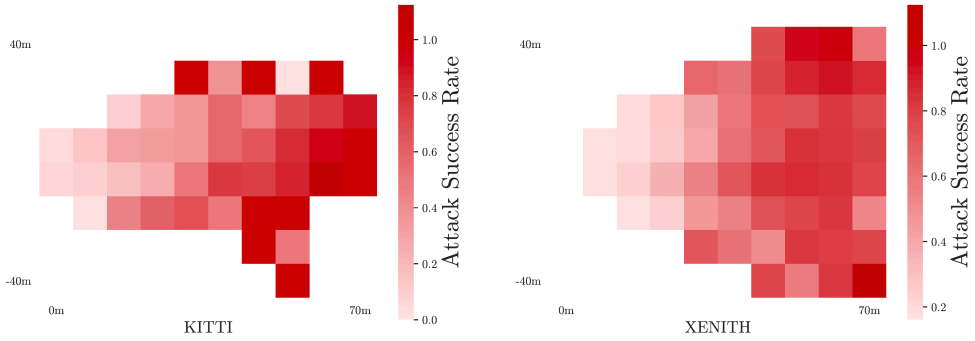

Figure 2: Visualization of the attack success rate across different locations in bird's eye view. The ego vehicle is at $(0, 0)$ with x, y being the longitudinal and lateral directions. Attacks are stronger on host vehicles that are more distant from the ego.

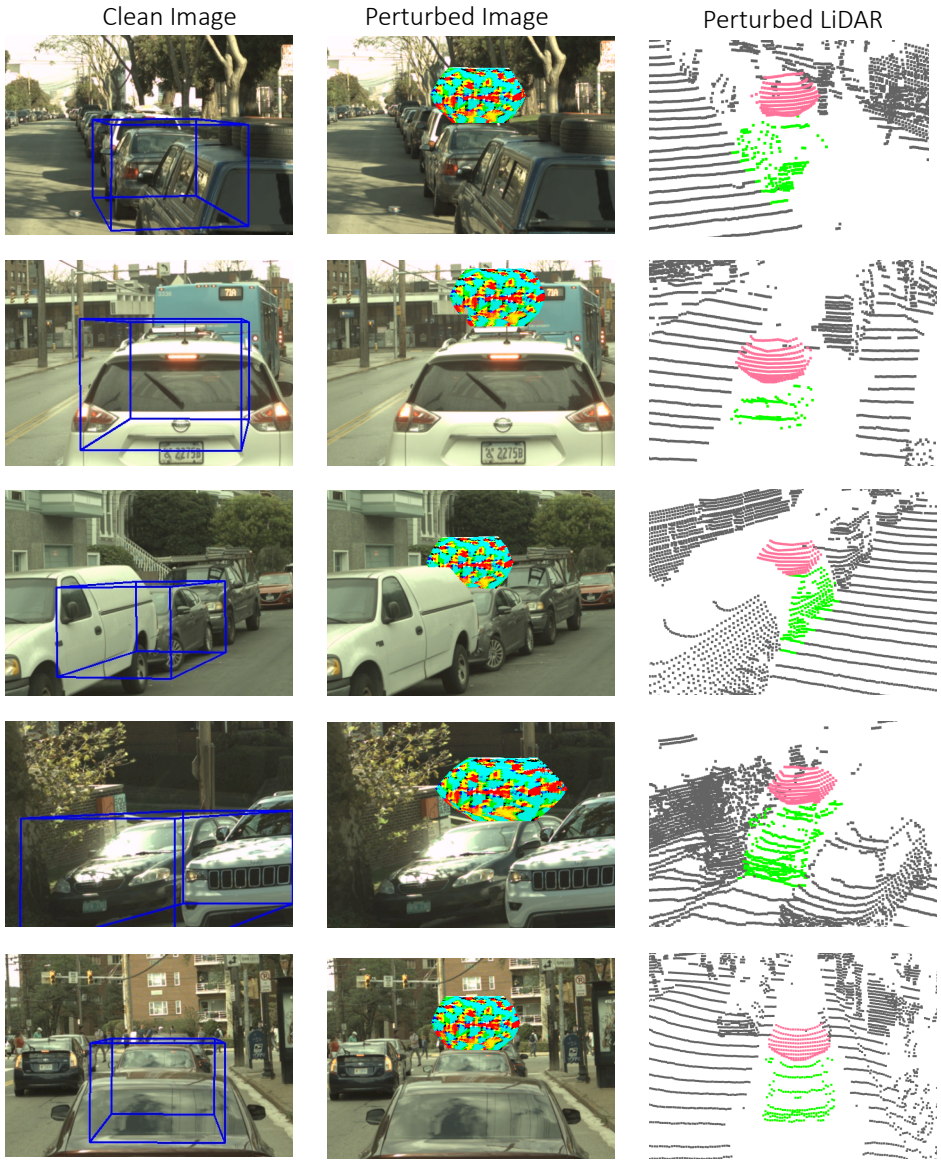

Figure 3: Additional qualitative examples on XENITH.

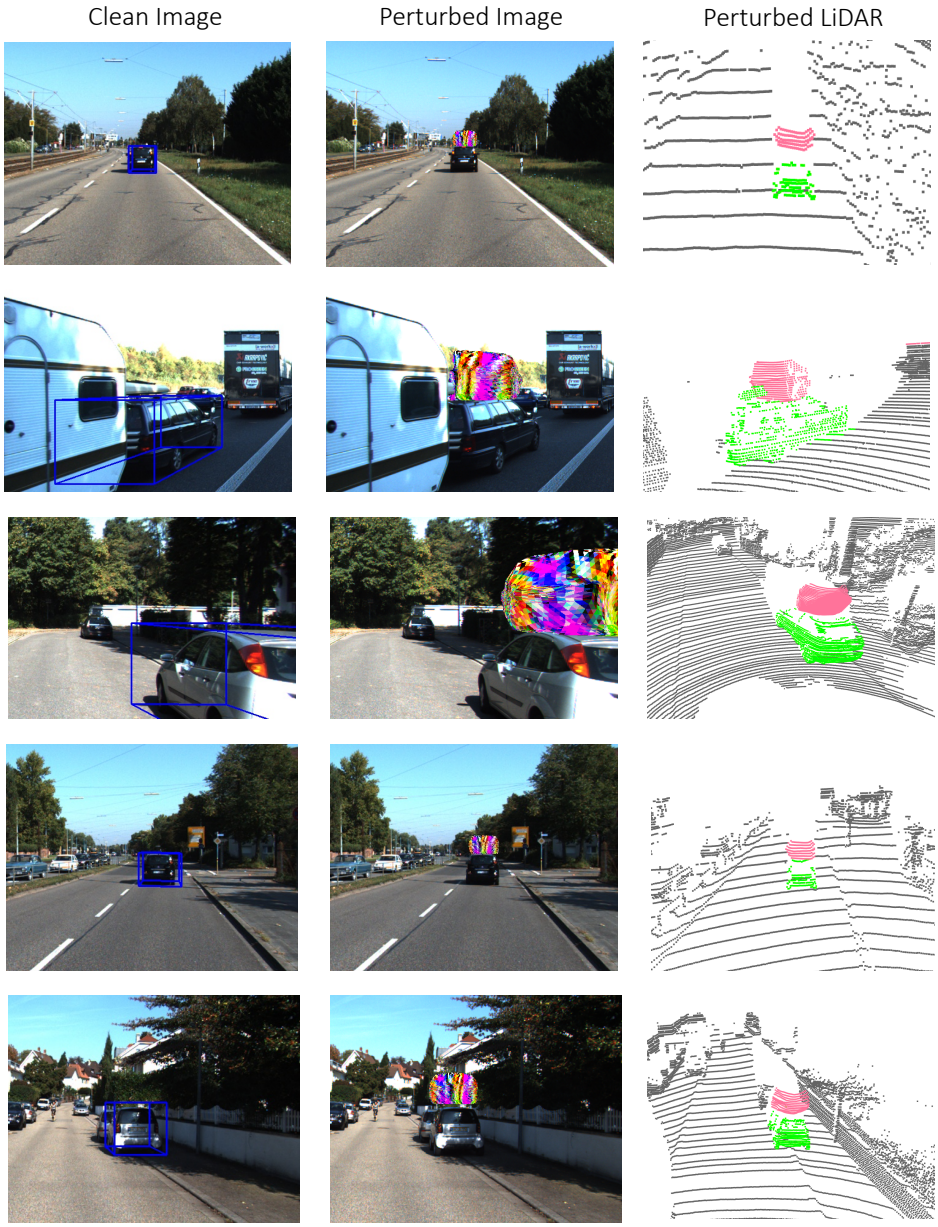

Figure 4: Additional qualitative examples on KITTI.

References

[1] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun. Multi-task multi-sensor fusion for 3D object detection. In *CVPR*, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Y. Chen, B. Yang, M. Liang, and R. Urtasun. Learning joint 2D-3D representations for depth completion. In *ICCV*, 2019.